УДК 004.75

Нархов С. А., Блеканов И. С.

Индекс для обработки и хранения гиперссылочных структур крупных Веб-сегментов

1. Введение. В настоящее время все больше организаций заботятся о качестве собственных веб-ресурсов в сети Интернет, а именно: мониторят и увеличивают рейтинги в поисковых системах; увеличивают индекс цитирования в Вебе; увеличивают вебометрический рейтинг [1] (в случае научно-общеобразовательных учреждений); улучшают коммуникабельность гиперссылочной структуры веб-ресурсов; улучшают эргономические свойства веб-ресурсов (удобочитаемость шрифтов, удобство глубины экрана, продолжительность пребывания пользователя на сайте, доступности целевого контента, эффективности брендинга в визуальном оформлении сайта и т. п.).

Исследованиями количественных и качественных характеристик гиперссылочной структуры занимается активно-развивающееся научное направление — вебометрика [2, 3]. Такие исследования веб-ресурсов называются вебометрическими. Местом апробации различных вебометрических методов, алгоритмов и программно-аналитических инструментов является университетский Веб [2, 4, 5], который включает крупные веб-сегменты, состоящие из сайтов научно-образовательных учреждений.

Важной особенностью сайтов университетского Веба является сложная разветвленная гиперссылочная структура большой размерности. В среднем, подобные веб-структуры ссылок содержат более 50 000 узлов и более 2 000 000 связей между ними [2, 6]. В связи с этим одной из актуальных задач вебометрики является задача индексирования гиперссылочных структур больших Веб-сегментов, решение которой в целом позволило бы существенно ускорить различ-

Нархов Семён Александрович – студент, Санкт-Петербургский государственный университет; e-mail: semennarhov@gmail.com, тел.: +7(921)345-52-35

Блеканов Иван Станиславович – доцент, Санкт-Петербургский государственный университет; e-mail: i.blekanov@gmail.com, тел.: +7(921)339-53-43

Работа выполнена при финансовой поддержке РФФИ, грант № 15-01-06105, и СПбГУ, НИР № 0.15.182.2015

ные вебометрические методы и алгоритмы анализа крупных сайтов, а также систематизировать хранение таких структур.

В данной работе в качестве решения поставленной выше задачи разработана индексная структура, основанная на использовании принципов инвертированного индекса.

2. Теоретическая часть. Идея разработанной индексной структуры схожа с идеей реализации прямого и обратного индекса для полнотекстового поиска [7]. Индекс состоит из следующих частей:

1. Словарь, содержащий множество страниц исследуемого вебсегмента, которым при обработке присваиваются уникальные целочисленные идентификаторы. Ключом в словаре является хэш от URL-адреса, а значением — идентификатор для этого адреса. Данная структура позволяет осуществлять поиск и добавление идентификатора для заданного URL-адреса страницы в среднем за $O(1)$. Для эффективного поиска (за время $O(1)$) URL-адреса любой страницы по идентификатору вместе со словарем в индексе используется специальный массив, в котором ключом является идентификатор, а значением — адрес страницы. Эти структуры расходуют $O(n)$ памяти, где n — число индексируемых страниц.
2. Множества входящих и исходящих ссылок для каждой страницы. Множество исходящих ссылок заранее известно, а для нахождения множества входящих ссылок выполняется следующая процедура: для каждой веб-страницы T пробегаем по всем исходящим ссылкам и для каждой исходящей ссылки L добавляем для страницы L входящую ссылку T. Данные множества ссылок хранятся в массивах переменной длины.

Кроме того, в индексе реализована возможность мониторинга гиперссылочной структуры исследуемого сайта во времени, которая основана на хранении в базе данных различных сессий индексирования данного сайта.

Теоретическая оценка времени построения разработанного индекса составляет $O(n)$.

3. Эксперимент. Ставился эксперимент, в котором разработанный индекс для обработки и хранения гиперссылочных структур апробировался на крупных сайтах научно-образовательных учреждений. В качестве сайтов из мирового вебометрического рейтинга

вузов [1] были выбраны два сайта российских университетов с наивысшим рейтингом:

1. Сайт Московского государственного университета.
2. Сайт Санкт-Петербургского государственного университета.

Для сбора и выявления гиперссылочной структуры веб-ресурсов использовался прототип программно-аналитической системы для вебометрических исследований, основанной на обобщенном ядре поискового робота и успешно апробированной в исследованиях [4, 5].

Эксперимент условно был разделен на три части. В первой части оценивалось время построения индекса для сайтов МГУ и СПбГУ. Во второй — время работы алгоритма Косарайю (алгоритм нахождения компонент сильной связности графа), использующего данные из построенного индекса. В третьей — производилась оценка скорости нахождения изменений в гиперссылочной структуре сайта МГУ (информация об удаленных и добавленных ссылках), собранной программным комплексом в разные моменты времени (в первой сессии сайт был проиндексирован полностью, во второй — проиндексирована лишь некоторая его часть, чтобы гарантировать наличие изменений в гиперссылочной структуре сайта).

4. Результаты эксперимента. Результаты оценки времени построения индексов для сайтов МГУ и СПбГУ приведены в таблице 1.

Таблица 1. Оценка времени построения индекса и работы алгоритма Косарайю

Параметры оценки	**Сайт МГУ**	**Сайт СПбГУ**
Количество обработанных индексом веб-страниц	45 379	127 682
Количество обработанных индексом ссылок	3 330 258	16 713 138
Индексирование исходящих ссылок, сек.	31,4	101,4
Индексирование входящих ссылок, сек.	0,71	3,7
Время выполнения алгоритма Косарайю, сек.	0,051	0,12

Результаты анализа изменения гиперссылочной структуры сайта МГУ между двумя сессиями сбора приведены в таблице 2.

Таблица 2. Оценка времени нахождения изменений в двух сессиях сайта МГУ

Параметры оценки	**Сессия сбора №1**	**Сессия сбора №2**
Количество веб-страниц	45 379	41 314
Количество ссылок	3 330 258	3 120 322
Количество добавленных ссылок		641 563
Количество удаленных ссылок		851 499
Поиск добавленных и удаленных ссылок, сек.		5,18

5. Заключение. Таким образом, результаты эксперимента показывают, что разработанный индекс имеет высокую скорость построения. Существенное различие скорости индексирования исходящих и скорости индексирования входящих ссылок (таблица 1) объясняется тем, что при индексировании исходящих ссылок приходится обращаться к базе данных, хранящей предварительно собранную программным комплексом веб-структуру сайта, в то время, как при построении индекса для входящих ссылок используется уже построенный индекс исходящих ссылок. Созданная авторами индексная структура также показала высокую скорость взаимодействия с алгоритмом Косарайю, что в результате обеспечило быстрый поиск компонент сильной связности (таблица 1).

Процедура получение информации из индекса по двум сессиям (таблица 2), хранящим состояние гиперссылочной структуры сайта МГУ, для определение разницы между ними, работает несколько медленнее, чем при взаимодействии индекса с алгоритмом Касарайю. Это связанно с загрузкой сессий из базы данных. Однако, это несущественно влияет на анализ гиперссылочных структур крупных сегментов веба.

Литература

1. Ranking Web of Universities (Main page). [Электронный ресурс]: URL:http://www.webometrics.info (дата обращения: 2.03.15).

2. Печников А. Вебометрические исследования Web-сайтов университетов России // Информационные технологии. 2008. № 11. С. 74–78.

3. Pechnikov A., Nwohiri A. Webometric analysis of Nigerian university websites [Электронный ресурс] // Webology. Vol. 9, Nom. 1, 2012. URL:http://www.webology.org/2012/v9n1/a95.html (дата обращения: 02.03.2015).

4. Блеканов И. С., Максимов А. Ю. Вебометрические исследования сегмента университетского Веба с помощью поискового робота // Процессы управления и устойчивость: Труды 44-й международной научной конференции аспирантов и студентов / под ред. Н. В. Смирнова, Т. Е. Смирновой. СПб.: Издат. Дом С.-Петерб. гос. ун-та, 2013. С. 403–408.

5. Blekanov I. S., Sergeev S. L., Maksimov A. I., Analysis of the topology of large Web segments using Broder's bow-tie model // Life Science Journal. 2014. Vol. 11. P. 258–261.

6. Блеканов И. С., Сергеев С. Л., Максимов А. Ю. Веб-краулер как инструмент для вебометрических исследований на примере анализа Веб-пространства СПбГУ // Materialy IX mezinarodni vedecko-prakticka konference «Moderni vymozenosti vedy — 2013». Praha, Czech Republic: Publishing House Education and Science, 2013. Vol. 71. P. 66–69.

7. Manning C., Raghavan P., Schutze H. An Introduction to Information Retrieval. England: Cambridge University Press. 2009. P. 6–10.